WHAT IS CLAIMED IS:

1. A method of constructing a trainable semantic vector representative of a data point in a semantic space, the method comprising the steps:

constructing a table for storing information indicative of a relationship between predetermined data points and predetermined categories corresponding to dimensions in the semantic space;

determining the significance of a selected data point with respect to each of the predetermined categories;

constructing a trainable semantic vector for the selected data point, wherein the trainable semantic vector has dimensions equal to the number of predetermined categories and represents the strength of the selected data point with respect to the predetermined categories.

2. The method of Claim 1, wherein the relative strength of the data points corresponds to the number of times each data point occurs in each category.

3. The method of Claim 1, wherein the data points correspond to words or other character strings occurring in a document.

4. The method of Claim 1, wherein the predetermined data points are contained within predetermined datasets.

5. The method of Claim 4, wherein the step of determining comprises the steps:

determining, for each category, the percentage of the predetermined datasets that contain the selected data point; and

determining the probability distribution of the selected data point's occurrences in the predetermined datasets across all categories.

6. The method of Claim 5 further comprising a step of minimizing the number of dimensions in the semantic representation of the selected data point.

7. The method of Claim 6, wherein the step of minimizing comprises the steps:

sorting the values of each dimension in the semantic representation of the selected data point in decreasing order;

determining a minimum number of dimensions for the semantic representation of the selected data point based on the sorted values; and

discarding all dimensions below the minimum number of dimensions.

8. The method of Claim 7, wherein the minimum number of dimensions is determined when at least 90% of the total mass of the semantic representation of the selected data point has been reached.

9. The method of Claim 6, further comprising a step of normalizing the value of the percentage of data points occurring in each category.

10. The method of Claim 9, further comprising a step of determining a weighted average of the normalized percentage of data points occurring in each category and the probability distribution of a data point's occurrence across all categories for each category.

11. The method of Claim 10, wherein the step of weighting is performed based on the formula

$$\alpha(v) + (1- \alpha)(u),$$

wherein α is a predetermined weighting factor, u is the normalized percentage of data points occurring in each category, and v is the probability distribution of a data point's occurrence across all categories.

12. The method of Claim 11, wherein the predetermined weighting factor has a value of about 0.75.

13. A method of constructing a trainable semantic vector representative of a data point contained within predetermined datasets in a semantic space, the method comprising the steps:

clustering the predetermined datasets into a plurality of unspecified clusters;

defining a plurality of categories such that each category corresponds to one of the plurality of unspecified clusters;

assigning each predetermined dataset to the category corresponding to the cluster to which the dataset belongs;

constructing a table for storing information indicative of a relationship between predetermined data points contained within the predetermined datasets and said plurality of categories, wherein each category corresponds to a dimension in the semantic space;

determining the significance of a selected data point with respect to each of the plurality of categories;

constructing a trainable semantic vector for the selected data point, wherein the trainable semantic vector has dimensions equal to the number of predetermined categories and represents the strength of the selected data point with respect to the predetermined categories.

14. A method of producing a semantic representation of a dataset in a semantic space comprising the steps:

constructing a table for storing information indicative of a relationship between predetermined data points within the dataset and predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the predetermined categories;

constructing a trainable semantic vector for each data point, wherein each trainable semantic vector has dimensions equal to the number of predetermined categories and represents the relative strength of its corresponding data point with respect to each of the predetermined categories; and

combining the trainable semantic vectors for the data points in the dataset to form the semantic representation of the dataset.

15. The method of Claim 14, wherein the semantic representation of the dataset is in the form of a vector having dimensions equal to the predetermined number of categories.

16. The method of Claim 14, wherein the dataset corresponds to a document, and each data point corresponds to a word or character string occurring in the document.

17. The method of Claim 14, further comprising a step of scaling the semantic representation of the dataset using a vote vector having entries that store a vote value for each dimension of the semantic representation of the dataset.

18. The method of Claim 17, wherein each vote value is at least 10.

19. The method of Claim 17, wherein each vote value is representative of the number of data points whose corresponding TSV dimension is greater than a predetermined minimum value.

20. The method of Claim 19, wherein the predetermined minimum value is about 0.5.

21. The method of Claim 14, further comprising a step of minimizing the number of dimensions in the semantic representation of the dataset.

22. The method of Claim 21, wherein the step of minimizing comprises the steps:

sorting the values of each dimension in the semantic representation of the dataset in decreasing order;

determining a minimum number of dimensions for the semantic representation of the dataset based on the sorted values; and

discarding all dimensions below the minimum number of dimensions.

23. The method of Claim 22, wherein the minimum number of dimensions is determined when at least 90% of the total mass of the semantic representation of the dataset has been reached.

24. The method of Claim 22, wherein the step of determining a minimum number of dimensions comprises the steps:

calculating the first derivative and second derivative of the semantic representation of the dataset at prescribed dimensions;

comparing the first derivative and second derivative to predetermined first and second pruning thresholds, respectively; and

identifying the minimum number of dimensions based on the step of comparing.

25. The method of Claim 24, wherein the first pruning threshold is about 0.05, and the second pruning threshold is about 0.005.

26. The method of Claim 24, wherein the derivatives are calculated in intervals of 10.

27. The method of Claim 24 wherein the step of identifying comprises the steps:

detecting a dimension at which the first derivative is lower than the first pruning threshold, and the second derivative is lower than the second pruning threshold;

doubling the value of the detected dimension;

comparing the doubled value of the detected dimension to a predetermined limit to determine a stop point corresponding to the lower value of the two; and

setting the minimum number of dimensions for the semantic representation of the dataset equal to the value of the stop point.

28. The method of Claim 27, wherein the predetermined limit is 1000.

29. A method of clustering data points from a dataset comprising the steps:

constructing a trainable semantic vector for each data point from the dataset in a multi-dimensional semantic space; and

applying a clustering process to the constructed trainable semantic vectors to identify similarities between groups of data points within the dataset.

30. The method of Claim 29, wherein the data points correspond to documents.

31. The method of Claim 29, wherein the step of applying a clustering process comprises the steps:

randomly distributing the data points among a predetermined number of clusters;

determining a cluster center for each cluster;

re-distributing the data points based on the determined cluster centers;

measuring an amount of change in each cluster; and

repeating the steps of determining, re-distributing, and measuring until a predetermined convergence factor has been reached.

32. The method of Claim 31, wherein:

the step of randomly distributing comprises a step of randomly assigning a fuzzy membership function to each data point; and

the step of re-distributing comprises the step of recalculating the fuzzy membership function for each data point.

33. The method of Claim 32, further comprising the step of making final cluster assignments based on the fuzzy membership functions.

34. The method of Claim 33, wherein each data point is assigned to zero or more clusters.

35. The method of Claim 31, wherein the step of randomly distributing comprises a step of randomly distributing an equal number of data points to each of the predetermined number of clusters.

36. The method of Claim 31, wherein the predetermined convergence factor is equal to about 0.0001.

37. The method of Claim 31, wherein the predetermined number of clusters is automatically determined based on the size of the dataset.

38. The method of Claim 31, wherein the predetermined number of clusters is input by a user.

39. The method of Claim 31, wherein the step of determining a cluster center comprises a step of constructing an average trainable semantic vector representative of an average value of all datasets within the cluster across all dimensions of the semantic space.

40. The method of Claim 39, wherein the step of re-distributing comprises a step of assigning the data points to clusters based on the distance from a data point to the nearest cluster center.

41. A method of classifying new datasets within a predetermined number of categories based on assignment of a plurality of sample datasets to each category, the method comprising the steps:

constructing a trainable semantic vector for each sample dataset relative to the predetermined categories in a multi-dimensional semantic space;

constructing a trainable semantic vector for each category based on the trainable semantic vectors for the sample datasets;

receiving a new dataset;

constructing a trainable semantic vector for the new dataset;

determining a distance between the trainable semantic vector for the new dataset and the trainable semantic vector of each category; and

classifying the new dataset within the category whose trainable semantic vector has the shortest distance to the trainable semantic vector of the new dataset.

42. The method of Claim 41 wherein the datasets correspond to documents.

43. The method of Claim 41 wherein the datasets correspond to email messages and the categories correspond to frequently asked questions with substantially static responses.

44. The method of Claim 41, further comprising the steps:

detecting when a prescribed number of new datasets has been classified; and

updating the trainable semantic vectors for each of the categories.

45. The method of Claim 44, wherein the step of updating comprises the step of re-constructing trainable semantic vectors for each category based on the trainable semantic vectors for the sample datasets and the trainable semantic vectors for the new datasets added to each category.

46. A method of classifying new datasets within a predetermined number of categories based on assignment of a plurality of sample datasets to each category, the method comprising the steps:

constructing a trainable semantic vector for each sample dataset relative to the predetermined categories in a multi-dimensional semantic space;

receiving a new dataset;

constructing a trainable semantic vector for the new dataset;

identifying a select number of sample datasets whose trainable semantic vectors are closest in distance to the trainable semantic vector for the new dataset; and

classifying the new dataset in the category containing the greatest number of the select sample datasets.

47. The method of Claim 46 wherein the datasets correspond to documents.

48. The method of Claim 46 wherein the datasets correspond to email messages and the categories correspond to frequently asked questions with substantially static responses.

49. The method of Claim 46, further comprising the steps:

detecting when a prescribed number of new datasets has been classified; and

adding the new datasets to the set of sample datasets.

50. A method of classifying new datasets within a predetermined number of categories, the method comprising the steps:

receiving a new dataset;

constructing a trainable semantic vector for the new dataset, where the dimensions of the trainable semantic vector correspond to the predetermined number of categories;

classifying the dataset in the category whose corresponding dimension in the trainable semantic vector has the largest value.

51. The method of Claim 50 wherein the datasets correspond to documents.

52. The method of Claim 50 wherein the datasets correspond to email messages and the categories correspond to frequently asked questions with substantially static responses.

53. A method of searching for datasets within a collection of datasets assigned to predetermined categories, the method comprising the steps:

constructing a trainable semantic vector for each dataset;

receiving a query containing information indicative of desired datasets;

constructing a trainable semantic vector for the query;

comparing the trainable semantic vector for the query to the trainable semantic vector of each dataset; and

selecting datasets whose trainable semantic vectors are closest to the trainable semantic vector for the query.

54. The method of Claim 53, wherein the datasets correspond to documents and the query is a natural language query.

55. The method of Claim 53, further comprising the steps:

performing a second search for datasets within the collection of datasets, wherein the second search using a method other than trainable semantic vectors;

combining the two search results to obtain a combined weighted score for each dataset in either of the two search results;

selecting datasets whose combined weighted score is largest.

56. The method of Claim 53, further comprising a step of clustering the selected datasets in real time.

57. A method of expanding a dataset, the method comprising the steps:

constructing a trainable semantic vector for the dataset;

comparing the trainable semantic vector for the dataset to the trainable semantic vectors of each of the data points in a semantic lexicon;

selecting data points whose trainable semantic vectors are closest to the trainable semantic vector for the dataset;

adding said selected data points to said dataset.

58. The method of Claim 57, wherein the dataset is a document and the data points are words.

59. The method of Claim 57, wherein the dataset is a natural language query in a search system and the data points are words.

60. A system for constructing a trainable semantic vector representative of a data point in a semantic space, the system comprising:

a computer configured to:

construct a table for storing information indicative of a relationship between predetermined data points and predetermined categories corresponding to dimensions in the semantic space;

determine the significance of a selected data point with respect to each of said predetermined categories;

construct a trainable semantic vector for said selected data point, wherein the trainable semantic vector has dimensions equal to the number of predetermined categories and represents the strength of the selected data point with respect to the predetermined categories.

61. A system for producing a semantic representation of a dataset in a semantic space, the system comprising:

a computer configured to:

construct a table for storing information indicative of a relationship between predetermined data points within the dataset and predetermined categories corresponding to dimensions in the semantic space;

determine the significance of each data point with respect to said predetermined categories;

construct a trainable semantic vector for each data point, wherein each said trainable semantic vector has dimensions equal to the number of said predetermined categories and represents the relative strength of its corresponding data point with respect to each of said predetermined categories; and

combine the trainable semantic vectors for the data points in said dataset to form the semantic representation of said dataset.

62. A system for clustering data points from a dataset comprising:

a computer configured to:

construct a trainable semantic vector for each data point from the dataset in a multi-dimensional semantic space; and

apply a clustering process to the constructed trainable semantic vectors to identify similarities between groups of data points within the dataset.

63. A system for classifying new datasets within a predetermined number of categories based on assignment of a plurality of sample datasets to each category, the system comprising:

a computer configure to:

construct a trainable semantic vector for each sample dataset relative to the predetermined categories in a multi-dimensional semantic space;

construct a trainable semantic vector for each category based on the trainable semantic vectors for the sample datasets;

receive a new dataset;

construct a trainable semantic vector for the new dataset;

determine a distance between the trainable semantic vector for the new dataset and the trainable semantic vector of each category; and

classify the new dataset within the category whose trainable semantic vector has the shortest distance to the trainable semantic vector of the new dataset.

64. A system for classifying new datasets within a predetermined number of categories based on assignment of a plurality of sample datasets to each category, the system comprising:

a computer configured to:

construct a trainable semantic vector for each sample dataset relative to the predetermined categories in a multi-dimensional semantic space;

receive a new dataset;

construct a trainable semantic vector for the new dataset;

identify a select number of sample datasets whose trainable semantic vectors are closest in distance to the trainable semantic vector for the new dataset; and

classify the new dataset in the category containing the greatest number of the select sample datasets.

65. A system for searching datasets within a collection of datasets assigned to predetermined categories, the system comprising:

a computer configured to:

construct a trainable semantic vector for each dataset;

receive a query containing information indicative of desired datasets;

construct a trainable semantic vector for the query;

compare the trainable semantic vector for the query to the trainable semantic vector of each dataset; and

select datasets whose trainable semantic vectors are closest to the trainable semantic vector for the query.

66. A computer-readable medium carrying one or more sequences of instructions for constructing a trainable semantic vector representative of a data point in a semantic space, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

constructing a table for storing information indicative of a relationship between predetermined data points and predetermined categories corresponding to dimensions in the semantic space;

determining the significance of a selected data point with respect to each of the predetermined categories;

constructing a trainable semantic vector for the selected data point, wherein the trainable semantic vector has dimensions equal to the number of predetermined categories and represents the strength of the selected data point with respect to the predetermined categories.

67. A computer-readable medium carrying one or more sequences of instructions for producing a semantic representation of a dataset in a semantic space, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

constructing a table for storing information indicative of a relationship between predetermined data points and predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the predetermined categories;

constructing a trainable semantic vector for each data point, wherein each trainable semantic vector has dimensions equal to the number of predetermined categories and represents the relative strength of its corresponding data point with respect to each of the predetermined categories; and

combining the trainable semantic vectors for the data points in the dataset to form the semantic representation of the dataset.

68. A computer-readable medium carrying one or more sequences of instructions for clustering data points from a dataset, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

constructing a trainable semantic vector for each data point from the dataset in a multi-dimensional semantic space; and

applying a clustering process to the constructed trainable semantic vectors to identify similarities between groups of data points within the dataset.

69. A computer-readable medium carrying one or more sequences of instructions for classifying new datasets within a predetermined number of categories based on assignment of a plurality of sample datasets to each category, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

constructing a trainable semantic vector for each sample dataset relative to the predetermined categories in a multi-dimensional semantic space;

constructing a trainable semantic vector for each category based on the trainable semantic vectors for the sample datasets;

receiving a new dataset;

constructing a trainable semantic vector for the new dataset;

determining a distance between the trainable semantic vector for the new dataset and the trainable semantic vector of each category; and

classifying the new dataset within the category whose trainable semantic vector has the shortest distance to the trainable semantic vector of the new dataset.

70. A computer-readable medium carrying one or more sequences of instructions for classifying new datasets within a predetermined number of categories based on assignment of a plurality of sample datasets to each category, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

constructing a trainable semantic vector for each sample dataset relative to the predetermined categories in a multi-dimensional semantic space;

receiving a new dataset;

constructing a trainable semantic vector for the new dataset;

identifying a select number of select datasets whose trainable semantic vectors are closest in distance to the trainable semantic vector for the new dataset; and

classifying the new dataset in the category containing the greatest number of the select datasets.

71. A computer-readable medium carrying one or more sequences of instructions for searching for datasets within a collection of datasets assigned to predetermined categories, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

constructing a trainable semantic vector for each dataset;

receiving a query containing information indicative of desired datasets;

constructing a trainable semantic vector for the query;

comparing the trainable semantic vector for the query to the trainable semantic vector of each dataset; and

selecting datasets whose trainable semantic vectors are closest to the trainable semantic vector for the query.